Supplemental materials: A principled approach for generating adversarial images under non-smooth dissimilarity metrics

Aram-Alexandre Pooladian*, **Chris Finlay, Tim Hoheisel, and Adam M Oberman**
McGill University, Department of Mathematics and Statistics

1 Proximal Operators

1.1 Motivation

We consider the following framework for proximal algorithms, namely a composite minimization problem

$$\min_{x\in\mathcal{E}} \Phi(x) := f(x) + g(x) \tag{1}$$

where $\mathcal{E}$ is an n-dimensional Euclidean space. We make the following assumptions:

- g is a non-degenerate, closed, and convex function over $\mathcal{E}$
- f is non-degenerate, closed function, with $\mathrm{dom}(f)$ convex, and has L-Lipschitz gradients over the interior of its domain
- $\mathrm{dom}(g) \subseteq \mathrm{int}(\mathrm{dom}(f))$
- the solution set, S, is non-empty.

Solving this composite problem with gradient descent is not advisable, since g is not necessarily differentiable. The best one can hope for is that g has a *subgradient* at $x \in \mathcal{E}$, defined as an element $v \in \mathcal{E}$ such that

$$g(y) \geq g(x) + \langle v, y - x\rangle \quad (y \in \mathcal{E}). \tag{2}$$

The collection of subgradients of g is called the *subdifferential* of g, denoted by $\partial g(\cdot)$. When a function is differentiable, the subdifferential is a singleton, namely $\partial f(x) = \{\nabla f(x)\}$. For an simple example of a subdifferentiable function, one can take the absolute value function;

$$\partial|\cdot|(x) = \begin{cases} +1 & \mathrm{sign}(x) > 0, \\ -1 & \mathrm{sign}(x) < 0, \\ [-1, 1] & x = 0. \end{cases}$$

Since Φ is a non-convex problem (because f is potentially not convex), our goal is to iteratively generate a sequence $\{x^{(k)}\}$ that converges to $x^* \in S$, where x^* is a *stationary point* i.e. $0 \in \partial\Phi(x^*)$. A characterization of these stationary points is the following fixed-point representation (we take $\lambda > 0$):

$$\begin{aligned} 0 \in \partial\Phi(x^*) &\iff 0 \in \nabla f(x^*) + \partial g(x^*) \\ &\iff x^* - \lambda\nabla f(x^*) \in x^* + \lambda\partial g(x^*) \\ &\iff x^* - \lambda\nabla f(x^*) \in (\mathrm{Id} + \lambda\partial g)(x^*) \\ &\iff x^* = (\mathrm{Id} + \lambda\partial g)^{-1}(x^* - \lambda\nabla f(x^*)) \end{aligned}$$

where $(\mathrm{Id} + \lambda\partial g)^{-1}(\cdot) =: \mathrm{Prox}_{\lambda g}(\cdot)$ is defined as the *proximal operator of g*

$$\mathrm{Prox}_{\lambda g}(x) := \mathrm{argmin}_{u\in\mathcal{E}} \left\{ g(u) + \frac{1}{2\lambda}\|x - u\|_2^2 \right\} \quad (\lambda > 0). \tag{3}$$

The first line in the equivalence chain uses addition of subdifferentiability, which is guaranteed by our assumptions, and the rest is algebraic manipulation. Thus, to generate a stationary point, it suffices to find a fixed point of the sequence generated in the following manner:

$$x^{(k+1)} = \mathrm{Prox}_{t_k g}(x^{(k)} - t_k\nabla f(x^{(k)})), \tag{4}$$

where $t_k > 0$ is some step size. The proximal operator exists for any convex function, but this is not a strict requirement.

1.2 Moreau Decomposition Theorem

The following is a result that is helpful for deriving proximal operators of ℓ_p norms.

Theorem 1 ***(Moreau Decomposition Theorem)*** *Let $f : \mathcal{E} \to \mathbb{R} \cap \{+\infty\}$ be closed, proper and convex. Then for $\lambda > 0$, the following holds:*

$$\mathrm{Prox}_{\lambda f}(x) + \lambda \mathrm{Prox}_{\lambda^{-1} f^*}(x/\lambda) = x,$$

where f^* is the conjugate function to f. While conjugate functions are outside the scope of this paper, we refer the interested reader to (Rockafellar and Wets, 2009) for more information. The following corollary follows

Corollary 1 *Let $f := \lambda\|\cdot\|_p$, with $f^* := \delta_{\mathbb{B}_q}$, where $\mathbb{B}_q$ is the unit ball for the dual norm to p, with $p^{-1}+q^{-1} = 1$. By the Moreau Decomposition Theorem,*

$$Prox_{\lambda\|\cdot\|_p}(x) = x - \lambda Proj_{\mathbb{B}_q}(x/\lambda).$$

1.3 Proximal operators for specific ℓ_p norms

In lieu of Cororollary 1, if we can perform efficient projections, then we have our proximal operators. For the proximal operator of the ℓ_∞ norm, we refer the reader to (Duchi et al., 2008). The runtime is $\mathcal{O}(n \log n)$; we have implemented a batch-wise version in our public repository. For the ℓ_2 norm, we perform a quick projection onto the ℓ_2 norm ball via normalization. Projections onto the ℓ_∞ norm ball is straight-forward, which gives the proximal operator for ℓ_1.

For the proximal operators for Total Variation and the ℓ_0 counting function, we refer the reader to (Beck, 2017) for a complete discussion. We remark that another interesting, non-differentiable dissimilarity metric is the Ordered-Weighted L1 (OWL) norm, which also has a proximal operator representation, see (Zeng and Figueiredo, 2014) for more information.

2 Adversarial Training

We briefly address details of our adversarial training methodology. On MNIST, we used the network described in (Papernot et al., 2015). In terms of adversarial training, we performed 40 steps of Projected Gradient Descent (PGD), and a constraint radius of 0.3 in the ℓ_∞ metric. On CIFAR10, we trained a ResNeXt-34 (Xie et al., 2016), and used 7 steps of PGD with radius 8/255 in the ℓ_∞ metric. The remaining hyperparameters are the same as those found in (Madry et al., 2017).

References

Beck, A. (2017). *First-order methods in optimization.*

Duchi, J., Shalev-Shwartz, S., Singer, Y., and Chandra, T. (2008). Efficient projections onto the l1-ball for learning in high dimensions. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pages 272–279, New York, NY, USA. ACM.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. (2017). Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*.

Papernot, N., McDaniel, P. D., Jha, S., Fredrikson, M., Celik, Z. B., and Swami, A. (2015). The limitations of deep learning in adversarial settings. *CoRR*, abs/1511.07528.

Rockafellar, R. T. and Wets, R. J.-B. (2009). *Variational analysis*, volume 317. Springer Science & Business Media.

Xie, S., Girshick, R. B., Dollár, P., Tu, Z., and He, K. (2016). Aggregated residual transformations for deep neural networks. *CoRR*, abs/1611.05431.

Zeng, X. and Figueiredo, M. A. T. (2014). The ordered weighted ℓ_1 norm: Atomic formulation, projections, and algorithms.